2020

An Interpretable Approach to Fake News Detection

Caio Brighenti
Colgate University, cbrighenti@colgate.edu

Bachelor Thesis

An Interpretable Approach to Fake News Detection

Caio Brighenti

Date: May 11, 2020

Advisors: Prof. Michael Hay and William Cipolli

Technical Report: COSC-TR-2020-03

Department of Computer Science
Colgate University
Hamilton, New York

Abstract

Misinformation has long been a tool for political influence, but it has taken a new form in the information age: fake news. After exploding into public consciousness during the 2016 United States presidential election, fake news has become a reality of political life around the world, featuring heavily in the 2017 German election and the 2018 Brazilian election. Fake news poses a significant threat to civic society, and is too easily produced and quickly disseminated to be resolved by manual fact-checking. As such, fake news detection has received significant attention by machine learning and natural language processing researchers in the last years. Previous work in this field has overly relied on deep learning approaches suffering from the black-box problem, rendering them unable to articulate precisely what properties separate fake news from real news. This paper contributes to the limited work on interpretable fake news detection by engineering text-based features, applying statistical tests, and fitting and interpreting logistic regression models. The results of this paper support previous findings that fake and real news are best differentiated by metrics capturing complexity and style, that fake headlines communicate far more than real ones, and that text-based approaches can effectively discern between real and fake news.

Acknowledgments

I would like to acknowledge and express gratitude to several people who provided invaluable feedback, advice, support, and without which this work would not have been possible.

Professor Cipolli for supporting this project from the outset, providing constant feedback over the semester, and for giving me the skills I needed to undertake this project in the classes I took with him.

Professor Hay for serving as a thesis advisor while leading a study group, and for providing incredibly useful feedback on both early drafts and presentations.

The Colgate University Division of Natural Sciences and Mathematics for funding the initial student-initiated research that kicked off this project in the Summer of 2019.

The Colgate University Computer Science Department for providing me with four great years of intellectual growth and learning.

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Hamilton, NY, May 7th, 2020

Caio Brighenti

(Caio Brighenti)

Contents

1. Introduction

1.1. Motivation

Propaganda has long been a tool of political influence, but in recent years it has taken a new online form: fake news. Fake news, once a buzzword on the internet, is now at the center of global politics. After the term gained prominence in Donald Trump's United States presidential campaign in 2016, it exploded into public consciousness, earning the distinction of Webster-Collins' "Word of the Year" in 2017.[5] As the current 2020 presidential race unfolds, fake news has returned to the center of attention, with major social media companies facing scrutiny of their misinformation policies. This phenomenon is also not a distinctly American problem–investigative reporting both during and after the 2018 Brazilian president election demonstrated that more than 40% of right-wing viral news articles shared on the popular messaging service WhatsApp were fake news favoring the eventual winner, Jair Bolsonaro.[2]

Fake news is not only politically significant but also dangerously tempting. Studies have shown false content propagates faster through social media than real content.[22] Blatantly false or exaggerated rhetoric can even lead to violent action, as demonstrated by the "Pizzagate" incident in which a man stormed a D.C. pizzeria with an AR-15, having been convinced by false and unverified information that a pedophile ring operated out of the restaurant's basement.[10] Fake news can also be dangerously easy to create. In 2019, a group of researchers at the Allen Institute for Artificial Intelligence published a text generation model able to produce fake news.[24] In a troubling conclusion, the researchers found that state-of-the-art fake news detection systems struggled more with identifying fake news produced by their systems than actual fake news. Fake news is thus easy to create, spreads quickly, and is hard to detect, a dangerous combination making it a serious threat to civic society.

1.2. Our Contribution

Given the danger that fake news poses, machine learning and natural language processing researchers have devoted significant attention to the problem of fake news classification. However, previous attempts at fake news classification overwhelmingly rely on highly complex models suffering from the black box problem. As a result, these models are unable to articulate *why* certain articles are classified as unreliable, and others as real. Automated fake news detection systems relying on these types of models are thus at risk of suffering from undiagnosable algorithmic bias or abuse, as their decision-making processes are unknowable.

Most importantly, the lack of clear justification behind the classification of content as fake or not presents a problem for effective debunking of fake news. While the prevalence of the "backfire effect"–whereby people may *strengthen* prior beliefs when presented with evidence to the contrary–is debated , researchers have shown "detailed debunking [is] associated with a stronger debunking effect than a nondetailed debunking."[16] Black box models, however, lack the interpretability needed to provide nuanced, detailed explanations for fake news classification.

In order to begin closing the black box gap, this research adopts an interpretable approach, with the overall objective of a producing a fake news classification model with comparable accuracy to state of the art models without compromising interpretability. Such an approach is not only positive for explaining predictions, but it can also contribute to a greater understanding of fake news *in general*. Identifying specific properties associated with and predictive of fake news provides data for psychologists, linguists, and other scholars more qualified to analyze how fake news sells misinformation. Additionally, identifying specific properties that are associated with fake news allows for clearer comparisons across studies, making it possible to determine what features are consistently signs of fake news.

The results of this paper make several contributions to the growing field of fake news detection. First, it demonstrates the effectiveness of interpretable, text-based models, with comparable accuracy to deep learning approaches on the same dataset. Second, this paper finds broad similarities with the findings of two papers that most closely resemble this paper's methodology. Third, despite the general agreements with prior work, the results of this paper show significant divergence at the feature level from comparable work, suggesting findings may not generalize across different datasets. Fourth and lastly, this paper contributes to the growing understanding of the textual characteristics of fake news, finding that the body of fake articles use more exclamation points and past-tense verbs, have a greater lexical diversity, and longer sentences. Furthermore, this paper finds that the headlines of fake articles also use more exclamation points and past-tense verbs, and differ primarily from reliable headlines in the number of unique words used.

1.3. Organization

Section 2 of this paper provides an extensive outline of previous work in fake news detection, describing how approaches tend to vary in levels of analysis, feature selection, and model type. Section 2 also demonstrates the need for interpretable approaches, highlighting two particularly relevant previous works that serve as bases for this paper's approach. Section 3 details the methodology employed in this paper, describing the dataset used, the feature engineering process, outlier removal, and the data analysis and modeling approach. Section 4 provides the results of the data analysis and modeling, including tables with significant results for Mood's Median Tests across all features, and plots showing the most impactful and important variables for fake news detection at the body-level and title-level. Finally, Section 5 summarizes the contributions of this paper, and offers several suggestions for further research informed by the results.

2. Related Work

Since the 2016 U.S. presidential election, fake news has been a frequent topic of natural language processing research. There are countless examples of papers approaching fake news classification or closely related problems from different angles. This section summarizes the prior work in fake news detection, clarifying the different categories of methods. In general, previous works in fake news detection differ in three major ways: 1) the scale of the predicted variable, 2) the information used as features, 3) and the type of model.

2.1. Levels of Analysis

With respect to scale, any fake news detection model falls into one of three levels of granularity: 1) claim level, 2) source level, and 3) article level. These levels of analysis describe the response variable being predicted.

Article level approaches have received the most attention in the work on fake news detection. This is logical, given that fake news tends to take the form of articles, peddling misinformation in the headline and body while posing as a legitimate source. Article-level approaches also benefit from a rich list of predictors to choose from, including not only the article's content but all relevant metadata. Shu et al., for instance, build an article-level model using linguistic and visual components of the article content, the social context around it—including information on the user that posted it, the post itself, responses to it, and the social network of the poster—, as well as spatiotemporal information capturing when and where the article and responses were to it were posted from.[18]

Claim level approaches attempt to determine whether specific short claims are true or intentionally misleading.[23] Given that claim-level approaches must make a judgement based on only a short amount of text, researchers often adopt a fact-checking strategy. This strategy, also known as "truth discovery," assumes that a sentence's truth claims can be gramatically isolated and checked against a database of established claims.[11] A natural application for claim-level models is social media, most commonly Twitter, where little is known about the author and only a very limited amount of text is available. However, claim-level approaches have serious limitations, often struggling with the complex sentences journalists or other

writers typically employ.[11] Additionally, they often rely on the existence of complete knowledge base, which must be constantly expanded and updated, clearly a difficult task.

Source level approaches attempt to classify whether a speaker or news source consistently publish fake news. The intuition behind these approaches is that speakers or sources that have published misinformation in the past are likely to continue to do so. An example of a source-level approach is the popular browser extension "BS Detector,"[19] which classifies articles on a fine-grained scale of veracity by checking the source's status in a database of news sources and their reliability. A source's history of misinformation can also be used as a predictor in claim-level or article level approaches. Kirilin and Strube, for instance, create *Speaker2Credit*, a metric of speaker credibility, and show how it can improve the performance of fake news detection models when used as an input.[11]

2.2. Feature Selection

Most relevant to this paper are studies leveraging features capturing textual properties of fake news. The two best examples are the work of Horne et al. and Gruppi et al., both of which employ features capturing the complexity, style, and psychology of fake news.[8][6] Specifically, the two use a highly overlapping feature set, with 29 features shared among both. Both papers leverage primarily features generated using the Linguistic Inquiry and Word Count tool, which processes texts into dozens of features capturing the complexity, grammar, style, and psychology of text. As will be described in Section 3, this paper adopts a similar approach to feature engineering.

In using predictors capturing linguistic, visual, spatiotemporal, and social components of article, the work of Shu et. al demonstrates the overwhelming number predictors available to researchers working in fake news detection.[18] In practice, this has resulted in a diversity of approaches with respect to feature engineering and feature selection. Tosik et. al, for instance, employ only hand-crafted features capturing the similarity between an article's title and text in a two-stage ensemble classifier modeling whether an article's body agrees with its headline.[20] Tripathi and Sharma demonstrate the effectiveness of parts of speech tagging—also known as grammatical tagging—in document classification problems, the general category of natural language processing that article-level fake news detection falls under. [21] Baly et al. employ a breadth of features to model factuality and bias of news sources, using features covering the content of articles, the source's Wikipedia and Twitter pages, the structure of the URL, and the source's web traffic.[3]

2.3. Model Types

Approaches that focus on interpretability are rare, but do exist. From the deep learning perspective, O'Brien et al. employ post-hoc variable importance to their text-based deep learning model, identifying the words that are most predictive of fake and real news.[14] Their work, however, does not interpret the results, but instead merely demonstrates the feasibility of the technique. Furthermore, this method reveals only information about *specific* words, as opposed to *types* of words. More applicable is the work of researchers who both use features that describe the semantic properties of the text in general, use interpretable models, and extensively document their results. As before, the most relevant examples are the work of Horne et al. Gruppi et al., both of which look for insights in their text-based features by applying statistical tests and non-neural network models.[8][6] Both works, however, limit their modeling to simply reporting the accuracy of certain combinations of predictors, and do not interpret model coefficients.

Aside from the few examples of interpretable approaches, researchers overwhelmingly choose to use complex deep neural networks. Ajao et. al, for instance, use a "hybrid of convolutional neural networks and long-short term recurrent neural network models" to classify Tweets as true or false based on their text content.[1] The dominance of deep learning approaches is visible in an extensive survey on fake news detection done by Oshikawa and Qian.[15] The pair's section on machine learning models dedicates a total of three sentences to "Non-Neural Network Models," compared to seven paragraphs focusing on neural networks.

While deep learning approaches can produce highly accurate models that consistently succeed in identifying misinformation, they also suffer from a lack of interpretability. This is often referred to as the black box problem, meaning that the inputs and outputs of these models are perfectly clear, but the steps that the model takes to reach the output are completely invisible. This has been identified as a limitation of the the work on fake news detection thus far.[18][14] Oshikawa and Qian, at the conclusion of their extensive survey, declare that "we need more logical explanation for fake news characteristics," highlighting the need for models that can teach us something about fake news.[15]

3. Methodology

This paper contributes to the small literature on interpretable fake news detection by following the methodology of Horne et al. and Gruppi et al., leveraging features that describe textual properties of fake news, and applying statistical tests and non-neural network models. In short, this paper builds upon the prior work of Horne et al. and Gruppi et al. by comparing the results of statistical tests between this paper and the two in order to verify consistency, and by interpreting the coefficients of the final models used. This section precisely details the methodology employed, discussing the dataset used, the feature engineering process, outlier removal, statistical testing, and models applied.

3.1. Dataset

There are many datasets freely available for fake news detection, but most suffer from a range of different limitations. Oshikawa and Qiang outline 12 requirements for a quality fake news dataset, expanding a 9-point list originally by Rubin et al.: 1. Availability of both truthful and deceptive instances; 2. Digital textual format acessibility; 3. Verifiability of ground truth; 4. Homogeneity in lengths; 5. Homogeneity of writing mattes; 6. Predefined timeframe; 7. The manner of news delivery; 8. Pragmatic concerns; 9. Language and culture; 10. Easy to create from raw data; 11. Fine-grained truthfulness; 12. Various sources or publishers.[15][17]

FakeNewsNet (FNN), the dataset used in this paper, is an article-level dataset that includes the title and body of each article (2),[1] each of which have been profesionally fact-checked and labeled as true or false by Politifact or Gossicop(1, 3).[18] FNN provides both political and celebrity news articles, but this paper chooses to use only the political articles in order to maintain a roughly consistent corpus (4,5,7). These articles are all in English, largely center around American politics, and come from a variety of sources (9, 12). Finally, there is little work needed to obtain the dataset, as FNN provides an API to quickly obtain the body and title of each article (10). The biggest limitation for FNN is that it lacks a fine-grained scale of truth, labeling only as binary true/false (11). However, given that it meets 9 out of

[1] (2) corresponds to the 2nd point on Oshikawa and Qian's 12-point list.

12 of Oshikawa and Qiang's criteria, and contains 969 observations, it is overall a good fit for this paper.

3.2. Feature Engineering

While FNN includes a host of metadata on each article, this paper utilizes only features engineered from the text and titles of each article. The objective of this paper is to reach new conclusions about the content of fake news, making certain metadata either irrelevant or problematic. For instance, the usage of website traffic to gauge veracity may increase accuracy,[18] but measures nothing about the actual content of an article. It should be obvious that websites with high traffic are capable of producing misinformation. Additionally, equating established sources with reliable sources can be problematic, failing to hold mainstream media accountable and preventing the growth of new, quality publishers.

This paper uses the FNN API to obtain the title and body of articles as well as a true/false label, then leverages a series of natural language processing tools to engineer features describing the text. Each feature is calculated for both the body and the title separately. Each feature falls under one of three categories: 1) complexity metrics, 2) stylistic and grammatical metrics, and 3) psychology metrics, in keeping with the feature set of Horne et al. and Gruppi et al.[2]

The complexity metrics are calculated in three different ways. Several of these metrics are indexes of textual complexity calculated using the "'quanteda"' package in R, relying on the relationship between the number of words, sentence length, and syllable counts. Others are variables describing the structure of verb-phrase and noun-phrase trees obtained for each sentence using the Stanford CoreNLP constituency parser.[13] Sentences with deeper constituency trees are naturally more structurally complex. The final complexity metric is a manually computed type-to-token ratio, where types are all the unique words in a document and tokens are the total words in that document, capturing the diversity of the vocabulary used.

The stylistic and grammatical features comprise a diverse range of different metrics, capturing different elements of the author's writing style and use of certain grammatical components. For instance, the stylistic metric "netspeak" captures the frequency of internet slang, while the grammatical metric "VBD" counts the frequency of past-tense verbs. Nearly all metrics in this category were computed using either the Stanford CoreNLP

[2]See Appendix A for a full list and description of each feature.

parts-of-speech tagger,[13] or the Linguistic Inquiry and Word Count (LIWC) tool.[9][3] The "stopwords" and "all_caps" metrics were manually computed. As these metrics rely on counting the incidence of words from dictionaries, each metric is normalized per 100 words, allowing them to be compared across differing document sizes without simply reflecting the document's length.

Overall, the feature set comprises 154 different features across the categories of complexity, style/grammar, and psychology. This paper's feature set shares a significant overlap with that of Horne et al. and Gruppi et al. Of the 80 unique features used in total by Horne et al. and Gruppi et al., 78 are included in this paper's feature set. The other 76 features, however, do not represent a significant departure from both papers. Both Horne et al. and Gruppi et al. leverage the same tools as this paper for feature engineering, but only use a subset of the available predictors. In order to avoid biasing results by arbitrarily selecting features that may intuitively appear to be related to fake news, this paper includes all available features and allows the testing and modeling processes to identify the most relevant ones.

The psychological features comprise a diverse range of dictionary-based metrics capturing the different psychological components within the text. In part, many of these relate to the author's emotion such as "sad" and "anger" which count the frequency of words relating to sadness and anger, respectively. Some are more complex, such as "reward," which counts the frequency of words indicating a focus on rewards, or "tentativeness," counting the frequency of words indicating tentativeness by the author.[4] As before, these features are normalized over 100 words to account for different document lengths.

Given titles tend to be short, 12 features had constant variance at the title level and were removed from title-level analysis. Additionally, the metrics capturing the distribution of the depth of constituency trees across the text calculated using CoreNLP are not relevant in cases where the text is a single sentence, and were thus not used for the statistical tests or modeling at the title level.

3.3. Outlier Removal

The complete training set thus consisted of 969 observations with 154 features each, constructed using the body and title of each article obtained by the FNN API, which queries the stored article URLs. However, given that web pages often change structure, move to different

[3]For more detailed descriptions of LIWC's metrics, consult Appendix A or the LIWC Operator's Manual available at https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf

[4]For a full description of how these features are computed, see the LIWC Operator's Manual available at https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf

addresses, or are removed from the internet entirely, many of the entries have changed since the dataset was initially compiled, and are now unavailable or in an incorrect format and must be removed from the dataset. To identify outliers, a baseline logistic regression model using all features was fit in order to identify overly influential observations with Cook's Distance four times larger than the mean. This approach cannot perfectly identify all outliers, but suggests that these observations are *potential* outliers. Each potential outlier was manually inspected, and flagged as either a true outlier or false positive. Other heuristics were used to identify potential outliers, such as filtering articles with an empty body or title, or that included words such as "error" or "unavailable" in the text. In all, 179 outliers were identified and removed, resulting in a final dataset of 790 articles.

3.4. Data Analysis and Modeling

After having completed all feature engineering and outlier removal, statistical tests were performed to identify group differences in each predictor between true and false articles. These tests were performed separately at the title and body level for each predictor. This approach reflects the work Horne et al. and Gruppi et al., allowing for the comparison of results between shared predictors. However, instead of applying an ANOVA test to each predictor, a Mood's Median hypothesis test was performed using the "'RVAdeMemoire"' package in R[7] as many of the predictors did not meet the normality assumption required by ANOVA and other traditional tests.

Two additional measures were taken to verify the statistical validity of this approach. Firstly, given the performance of 308 individual tests (once for the title and once for the body of each of the 154 predictors) raises the possibility of false positives induced through multiple testing, a p-value adjustment method was applied. Specifically, the Benjamini-Hochberg p-value adjustment was used, which punishes p-values according to their ranking in order to control the rate of false positives.[4]

Second, to verify that power of Mood's Median Test was not inflated with the sample size in this study, an experiment was conducted to verify the false positive rate of the test at this sample size. This experiment consisted of generating 1000 samples of 790 observations of random normally distributed data, with each observation having class of 0 or 1, to which a Mood's Median Test was applied. With 95% confidence, roughly 50 out of 1000 samples should result in significant tests with p-values below 0.05. This experiment resulted in 47 'significant' samples, performing nearly exactly as expected for a test with 95% confidence. This confirms that the results of the Mood's Median Tests are not likely to be products of a large sample size.

Before proceeding to model fitting, three data processing steps were taken to prepare the dataset. First, the full dataset of 790 observations was randomly split into a train and test set with 577 and 213 observations, respectively, following a typical 75:25 split. Second, to avoid skewed accuracy results due to class imbalance, the dataset was upsampled to have an equal proportion of negative and positive labels. The upsampling increased the number of negative articles to have a 374-374 split, as opposed to the original 374-220 split. After these two steps were performed, the dataset was prepared for model fitting. Third, in order to allow the coefficients to be comparable across metrics with different ranges, all features were normalized to have zero means and standard deviations of one.

Given the extensive number of predictors and the multicolinearity between many of them, a binomial Lasso regression—also known as logistic regression with L1 regularization—was applied to avoid overfitting by creating a more parsimonious model. The regularization parameter λ was selected through cross validation. The final λ selected was not the one that resulted in the lowest mean squared error, but rather the largest λ within 1 standard error from the 'ideal' λ. This was done in order to produce a more parsimonious model, due to the large number of predictors.

After fitting the Lasso models, the features reduced to zero were removed and final logistic regression models at the body level and title level were fit using only the preserved features. These models contained 42 and 45 features, respectively, and are the models used for final interpretation of results. An approach focused entirely on predictive accuracy would likely instead create a two-stage ensemble model, but as the overall objective is interpretation, maintaining the two separate models is preferable as it allows for better interpretation of the results.

The results of the models are shown in the form of plots capturing the most important and impactful variables. Variable importance is measured using the "caret" package in R.[12] The variable importance metric used is the AUC of each feature when used in a univariate model predicting the class in question. This gives a measure of each feature's individual predictive strength with a baseline of 0.5. The feature with the highest variable importance score has the highest individual predictive power across the entire dataset. Variable impact is measured using the coefficients of the final logistic regression model. While important variables have high predictive power across the entire dataset, impactful ones have the highest effect at the observation level when taking on a value significantly higher or lower than the mean for that feature.

4. Results

This section summarizes the results of the analysis, starting with the pairwise Mood's Median tests followed by the results of the modeling process. Results are compared with the work of Horne et al. and Gruppi et al. to highlight overlaps and disagreements between their results and the results of this paper.

4.1. Mood's Median Test

As described in Section 3, a Mood's Median Test was used to test for differences between real and fake news for each of the 154 features at both the title and body level. The table below includes the result of these tests, excluding tests with resulting adjusted p-values of > 0.05. For each predictor, the p-value of the test is given, along with a comparison of the groupwise medians to indicate which group is higher.

One of the core objectives of this paper is to determine whether the results of this paper are consistent with the prior work of Horne et al. and Gruppi et al. Given the significant overlap of features, if the results of both papers are applicable to fake news in general, the results of this paper should be consistent with the results of both. As such, the tables below include columns showing whether this paper's results agreed or disagreed with the results in Gruppi et al. and Horne et al.

Note that only features significant in this paper and one of Horne et al. or Gruppi et al. are included, and that the agreement column is concerned with the directionality of the relationship. For instance, the overall word count of articles was significantly different between fake and real news in all three cases, but Horne et al. and Gruppi et al. found that this feature was *higher* for fake news, while this paper found that this feature was *lower*.

Table 4.1.: Mood's Median Test Results for Article Body

Variable	Result	p-value	Gruppi et al.	Horne et al.
mu_sentence	Fake < Real	< 0.001	-	Disagree
mu_verb_phrase	Fake < Real	0.0126	-	Disagree
num_verb_phrase	Fake > Real	< 0.001	-	-
swc	Fake < Real	< 0.001	-	-
types	Fake > Real	< 0.001	-	-
tokens	Fake > Real	< 0.001	-	-
TTR	Fake < Real	< 0.001	-	Disagree
FOG	Fake < Real	0.0021	-	-
SMOG	Fake < Real	0.0054	Agree	-
FK	Fake < Real	< 0.001	Agree	Agree
CL	Fake > Real	0.0187	-	-
ARI	Fake < Real	< 0.001	-	-
all_caps	Fake < Real	< 0.001	Agree	Disagree
stopwords	Fake > Real	< 0.001	-	-
WC	Fake > Real	< 0.001	Disagree	Disagree
Authentic	Fake > Real	0.0367	-	-
Tone	Fake > Real	< 0.001	-	-
WPS	Fake < Real	< 0.001	Disagree	-
Dic	Fake > Real	0.0215	-	-
shehe	Fake < Real	< 0.001	Agree	Disagree
article	Fake < Real	< 0.001	Agree	-
prep	Fake < Real	< 0.001	-	-
auxverb	Fake < Real	0.0021	Agree	-
conj	Fake > Real	0.0015	-	-
verb	Fake < Real	0.0367	-	-
quant	Fake > Real	0.0064	Agree	-
posemo	Fake > Real	< 0.001	-	Disagree
negemo	Fake < Real	< 0.001	-	Agree
anger	Fake < Real	< 0.001	-	-
male	Fake < Real	< 0.001	-	-
cogproc	Fake > Real	< 0.001	-	-
cause	Fake > Real	0.0247	-	-

Table 4.1.: Mood's Median Test Results for Article Body *(continued)*

Variable	Result	p-value	Gruppi et al.	Horne et al.
differ	Fake > Real	0.0084	-	-
see	Fake < Real	< 0.001	-	-
hear	Fake < Real	0.0029	-	-
bio	Fake < Real	< 0.001	-	-
drives	Fake > Real	< 0.001	-	-
affiliation	Fake > Real	< 0.001	-	-
achieve	Fake > Real	< 0.001	-	-
focuspast	Fake < Real	< 0.001	-	-
focuspresent	Fake > Real	< 0.001	-	-
relativ	Fake < Real	0.0367	-	-
time	Fake < Real	< 0.001	-	-
work	Fake > Real	< 0.001	-	-
leisure	Fake < Real	< 0.001	-	-
money	Fake > Real	< 0.001	-	-
AllPunc	Fake > Real	0.0367	-	Disagree
Period	Fake > Real	< 0.001	-	-
Dash	Fake > Real	0.0021	Agree	-
Quote	Fake < Real	< 0.001	-	Agree
CC	Fake > Real	< 0.001	-	-
IN	Fake < Real	< 0.001	-	-
JJR	Fake > Real	0.0187	-	-
JJS	Fake > Real	0.0197	-	-
MD	Fake > Real	0.0034	-	-
NNPS	Fake > Real	0.0054	-	-
POS	Fake < Real	< 0.001	-	-
PRP$	Fake < Real	0.0034	-	Disagree
RBR	Fake > Real	< 0.001	-	-
RP	Fake < Real	< 0.001	-	-
VB	Fake > Real	< 0.001	-	-
VBD	Fake < Real	< 0.001	-	-
VBG	Fake < Real	< 0.001	-	-
VBN	Fake < Real	< 0.001	-	-

Table 4.1.: Mood's Median Test Results for Article Body *(continued)*

Variable	Result	p-value	Gruppi et al.	Horne et al.
VBP	Fake > Real	0.0014	-	-
NER	Fake < Real	< 0.001	-	-

Despite significant overlap in features used with two comparable papers, very few features significant in this paper's tests were also significant in the other two studies. Furthermore, when features were significant in both cases, they frequently disagreed, in many cases even demonstrating agreement for one study and disagreement in another. At the body level, only the Flesh-Kincaid readibility score (FK) was significant in all three and agreed in direction in all three. This prevalence of disagreement raises questions about the generalizability of the findings in Horne et al. and Gruppi et al.

Next, the results for the same tests and features with respect to the title of articles is shown below.

Table 4.2.: Mood's Median Test Results for Article Title

Variable	Result	p-value	Gruppi et al.	Horne et al.
types	Fake < Real	< 0.001	-	-
tokens	Fake < Real	< 0.001	-	-
TTR	Fake > Real	< 0.001	Disagree	-
SMOG	Fake < Real	< 0.001	-	-
all_caps	Fake > Real	< 0.001	Agree	Agree
stopwords	Fake < Real	< 0.001	-	Agree
WC	Fake < Real	< 0.001	Disagree	-
WPS	Fake < Real	< 0.001	Disagree	Disagree
Sixltr	Fake > Real	0.042	Disagree	-
function.	Fake < Real	< 0.001	-	-
prep	Fake < Real	< 0.001	-	-
verb	Fake < Real	< 0.001	-	-
social	Fake < Real	< 0.001	-	-
space	Fake < Real	0.0082	-	-
work	Fake > Real	0.0041	-	-
IN	Fake < Real	0.005	-	-
NER	Fake > Real	0.0169	-	-

Similarly to the results at the body level, there was little overlap between significant results in this paper and in the work of Horne et al. and Gruppi et al., with only six features significant in at least two. Aditionally, the pattern of disagreement continued, with only the frequency of capitalized letters and the frequency of stopwords–those being syntatically null words such as "the" or "which"–demonstrating agreement. This is further evidence that the results following this methodology may depend heavily on the dataset used, and might not generalize well to other fake news datasets.

4.2. Modeling

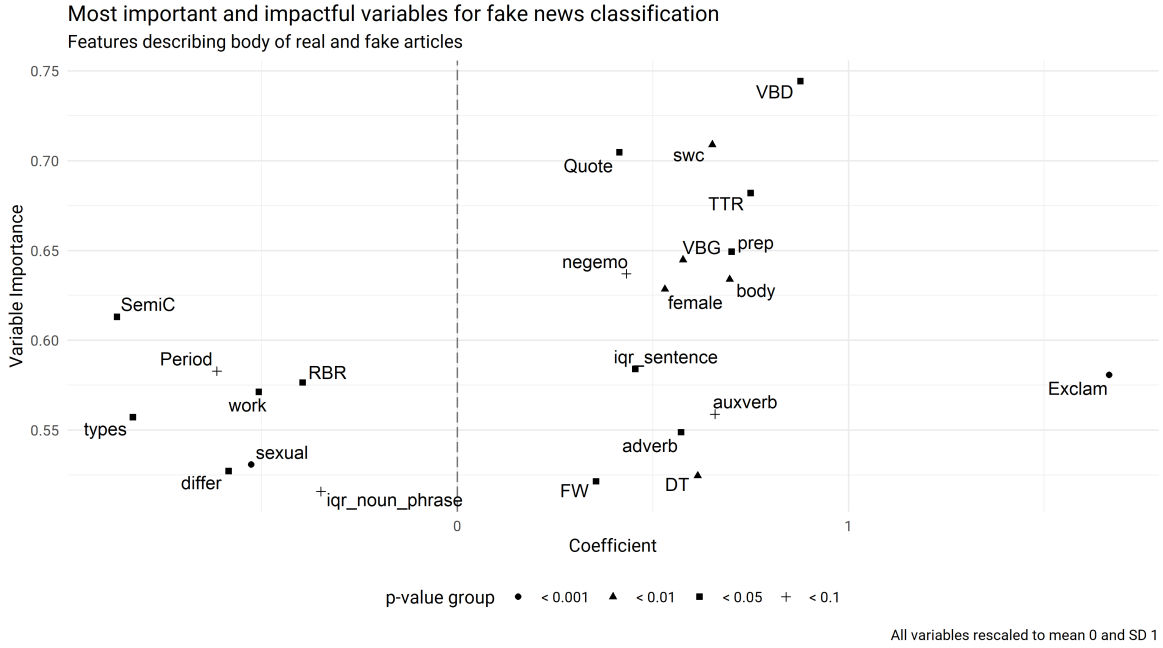
While the results of the Mood's Median Tests demonstrate where real and fake articles differ, on median, at the level of each feature, it does not provide a meaningful way to predict the veracity of an article based on these textual features. While we expect features with significant tests to be good predictors, this is not always the case. This section includes the actual results of modeling, highlighting the most important and impactful predictors of fake news, utilizing the coefficients and AUC-based variable importance. The overall predictive accuracy of each model is also included and compared to prior work with the FNN dataset, to demonstrate the efficacy of this paper's model.

The final body-level model included 42 features after the Lasso regression shrunk the coefficients of 112 features to zero. This model performs quite well, with an AUC of 0.96 on the train set, and a accuracy on the held-out test set of 79% relative to a 50% baseline, as well as sensitivity and specificity of 0.77 and 0.8, respectively, with a cutoff of 0.5. This is in line with the results of Horne et al. and Gruppi et al., which ranged from 70%-80% accuracy, and comparable to state-of-the-art models applied to FNN. The best approaches to FNN range from 87%-93%, which is an improvement from this paper's model, but is close enough to be comparable.

The most impactful and important features for the body-level model are shown in the following plot, with the coefficient magnitude shown on the x-axis and the AUC-based variable importance on the y-axis. Note that the baseline importance for a variable with no predictive power is 0.5, hence why the y-axis does not start at 0. Features with large variable importance scores have the most individual predictive power, while features with large coefficients having the largest effect on individual observations when these features take on values significantly away from the mean. Negative coefficients represent features indicative of real news, while positive ones are indicative of fake news. Note that only features with p-values of <0.1 are shown.[1]

[1]The complete coefficient table for the model is available in Appendix B.

Figure 4.1.: Variable Impact and Importance for Body-Level Model



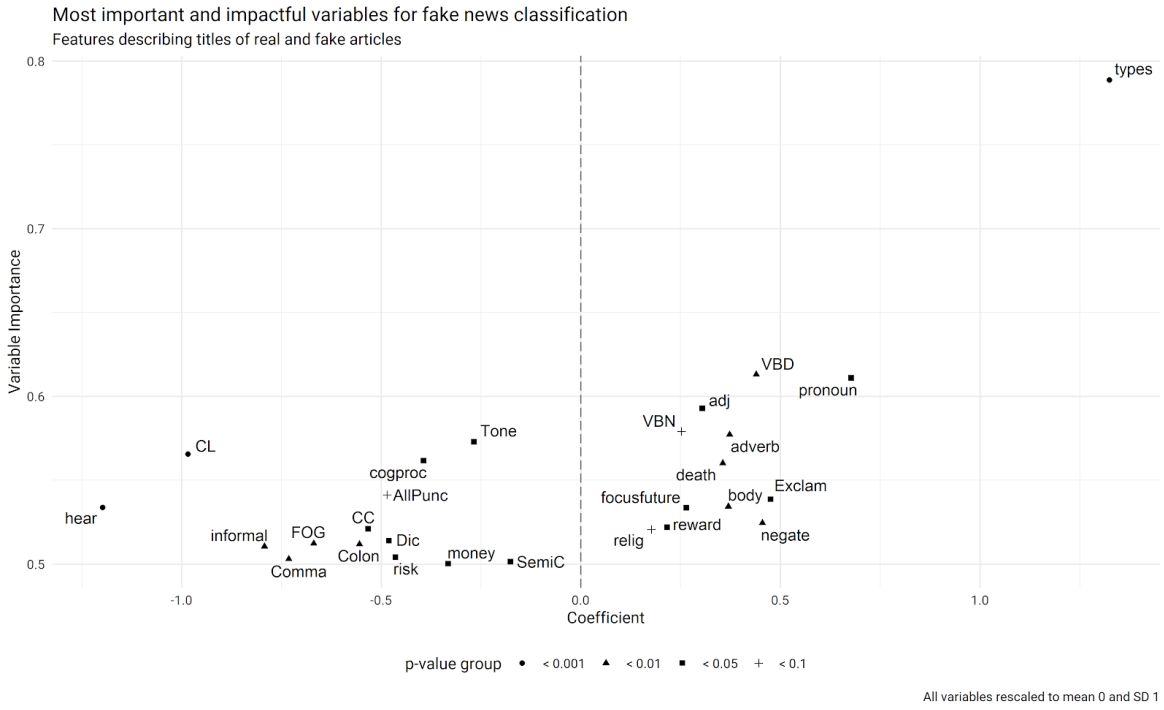
Many insights can be gleaned from each of the predictors shown in the above plot, but three groups stand out. First, the frequency of past-tense verbs (VBD), sentence word count (swc), type-to-token ratio (TTR), and frequency of quotes are the most important indicators of fake news across the dataset. Second, the frequency of exclamation points is by far the most impactful predictor of fake news. This suggests that exclamations marks may not be widespread among fake news, but when their frequency is significantly far from the mean, it strongly indicates the presence of fake news. Third, the frequency of semicolons and number of unique words are the most impactful predictors of real news.

The final title-level model included 45 features after the Lasso regression shrunk the coefficients of 109 features to zero. This model performs similarly to the body-level model, with an AUC of 0.94 on the train set, and an accuracy of 81% relative to a baseline of 50%, as well as a sensitivity and specificity of 0.76 and 0.88, respectively, using a cutoff of 0.5 selected using ROC curve analysis. Again, this accuracy is similar to prior work, in line with the work of Horne et al. and Gruppi et al., and comparable to state-of-the-art approaches to FNN.

A feature importance and impact plot was also created for the title-level model and is shown below. The interpretation for this plot is the same as before. Again, only features with

p-values lower than 0.1 are shown.[2]

Figure 4.2.: Variable Impact and Importance for Title-Level Model



The results at the title level paint a clearer picture than at the body level. The "types" feature, capturing the number of unique words in the title, is by far the most important *and* impactful predictor of fake news. This indicates titles containing more information and subsequently using more unique words are a strong indicator of fake news. Real titles, on the other hand, tend to use words related to hearing (hear) more frequently, and have higher Coleman-Liau complexity scores (CL), on average. Similarly to the body of fake news, fake titles use more past-tense verbs and exclamations. Additionally, fake titles use more negations and pronouns, on average.

4.3. Visualizing Features

Given that the features in this paper are quantitatively measuring properties of text, they can be difficult to intuitively understand. Using the results of the title-level model, this

[2]The complete coefficient table for this model is also available in Appendix B.

section briefly visualizes some examples of how these features are encoded and understood by the model by highlighting where they can be identified in examples of fake headlines. In the following title samples, yellow corresponds to pronouns, green to exclamation marks, blue to past-tense verbs, and red to negations.

> *BOMBSHELL !! Obama Paid FBI Informant Over $1 MILLION To Do It To Trump ! EVEN AFTER THE ELECTION !*

> *Michelle Obama: 'Florida Shooting Is Clearly Trump's Fault, These Shootings Are Happening Constantly Since He Became Our President. We Must Protect Our Children From This Tyrant"Do You Support Her ?*

> *Actress Sandra Bullock to Hillary Clinton if You Don't Like Our President You Can Leave and Never Come Back Again You Are One Jealous Woman Who is Nothing to Compare With Trump I Hope He Will Arrest You*

Throughout these fake headlines, it becomes clear how the presence of exclamation points, past-tense verbs, pronouns, and negations are interpreted by the model. Most significantly, these titles are great examples of why "types," the number of unique words, is such a strong predictor of fake news, as all three titles are long and contain a great deal of information.

5. Conclusion

This section summarizes the contributions of this paper, describing the effectiveness of an interpretable text-based model, discussing the similarities and divergences from comparable work, outlining the biggest takeways with respect to identifying fake news, and finally offering suggestions for future work in fake news detection.

5.1. Contributions

This paper set out to build an interpretable fake news detection model using only features describing the textual properties of article. One objective was to demonstrate that such a model can perform close to deep learning approaches, while simultaneously preserving the learning opportunity offered by interpreting results and human-understandable features. Given that the accuracy of both models were comparable to that of more advanced models on the same dataset (FNN), it is clear that one need not forgo accuracy entirely when focusing on interpretability. Deep learning approaches will certainly perform better, and have their applications, but the results of this paper suggest more attention should be paid to interpretable models as deep learning approaches dominate the field.

This paper's most significant findings relate to the similarities and dissimilarities with the prior work of Horne et al. and Gruppi et al. Despite using a highly overlapping feature set, the results of the statistical tests employed by this paper differed heavily from the results of both Horne et al. and Gruppi et al. If the results of the three works were applicable to fake news *in general* and not representations of the specific datasets used, we would expect to see significant agreement in these tests. This divergence thus calls into question whether textual indicators of fake news are consistent across different datasets.

While this paper's results may have differed significantly from that of Horne et al. and Gruppi et al. at the feature level, there were significant similarities more generally. At a high level, both Horne et al. and Gruppi et al. reach three main conclusions: 1) complexity and stylistic measures are the best differentiators of fake news, 2) the headlines of fake news attempt to include much more information than real headlines, and 3) text-based approaches are effective at identifying fake news with accuracy between 70%-80%. All

three general findings were confirmed by the results of this paper. The complexity and stylistic measures were most important, with almost no psychological metrics appearing in the model predictor plots at either the body or title level. Given the strength of "types" as a predictor of fake news at the title level, this paper's results strongly agree that fake headlines attempt to communicate more than real ones. Finally, as described above, this paper demonstrated the effectiveness of a text-based approach, with accuracies of 79%-81%.

This paper's final contribution relates to the precise features identified as being indicative of fake and real news. Considering only the observed results of this paper, several text features make for good predictors of fake news. Fake news uses more past-tense verbs, exclamations, and quotes at the body level, and have longer sentences and a more diverse vocabulary. Real news, on the other hand, uses semicolons more frequently, and have a greater number of unique words in total. Given that the type-to-token ratio is larger in fake news than real news, on average, while larger numbers of unique words (types) is indicative of real news, we can infer that the body of fake news has less words and content overall when compared to fake news, decreasing the denominator in the type-to-token ratio. At the title level, this paper identifies the number of unique words, frequent usage of pronouns, exclamation points, negations, and past-tense verbs as indicative of fake news. This suggests headlines of fake news focus more on the past (past-tense verbs), people (pronouns), are more emotive (exclamation marks), and fit more information in overall (types).

5.2. Future Work

The results of this paper suggest several different viable research paths for fake news detection. Firstly, it is obvious more works following the methodology put forth in this paper and in the work of Horne et al. and Gruppi et al. Given the significant divergence between the results across the three papers, there is a serious need for further work to identify which properties of fake news are truly significant and not spurious or a product of the specific datasets used. Further work in this methodology should apply the chosen models and tests to a variety of different datasets in order to identify what properties are consistent across fake news *in general*.

In keeping with expanding the coverage of datasets, future work should also include a longitudinal component. Fake news detection is an inherently adversarial objective, whereby researchers hope to defeat those that produce fake news, while the peddlers of misinformation attempt to improve their techniques and make their articles more believable. Studying whether properties of fake news have changed over time, perhaps adapting to published work in the field, would be important in identifying the consistency of fake news over time.

The results of this paper also demonstrate that the lack of interpretable studies is unjustified, given the strong performance of a simple logistic regression model on FNN. As the over-reliance on black box models has been identified as a significant limitation of prior work in this field, it would seem particularly prudent for more researchers to follow this direction. This is not to say that deep learning approaches should be abandoned, but rather that such a direction is oversaturated while little attention has been paid to interpretable approaches.

Finally, while an interpretable model can serve as a learning opportunity regarding fake news, offering specific textual properties distinguishing real and fake news, a computer science researcher is not qualified to fully interpret these properties. The overall research objective of identifying relevant textual properties of fake news and using them to better understand the phenomenon in general can only be done in collaboration with researchers from other fields, such as linguists and psychologists. Given the threat that fake news poses, this type of interdisciplinary work is critical, and should be pursued immediately. If understanding fake news is not treated as a serious research objective, the democratic process of the upcoming U.S. presidential election is in significant risk of being compromised.

A. Variable Descriptions

A.1. Complexity Metrics

Table A.1.: Complexity Metrics

Variable	Description
mu_sentence	Mean depth of sentence constituency trees
mu_verb_phrase	Mean depth of verb-phrase trees
mu_noun_phrase	Mean depth of noun-phrase trees
sd_sentence	Standard deviation of depth of sentence constituency trees
sd_verb_phrase	Standard deviation of depth of verb-phrase trees
sd_noun_phrase	Standard deviation of depth of noun-phrase trees
iqr_sentence	Interquantile range of depth of sentence constituency trees
iqr_verb_phrase	Interquantile range of depth of verb-phrase trees
iqr_noun_phrase	Interquantile range of depth of verb-phrase trees
num_verb_phrase	Number of verb-phrase trees
swc	Mean sentence word count
wlen	Mean word length
types	Number of unique words
tokens	number of total words
TTR	Type-token ration
FOG	Gunning's Fog Index
SMOG	Simple Measure of Gobbledygook
FK	Flesch-Kincaid Readability Score
CL	Coleman-Liau Index
ARI	Automated Readability Index

A.2. Stylistic Metrics

Table A.2.: Stylistic Metrics

Variable	Description
all_caps	Frequency of capitalized characters
stopwords	Frequency of stopwords
WC	Word count
WPS	Words per sentence
Sixltr	Frequency of of six+ letter words
Dic	Frequency of words present in LIWC dictionary
function	Frequency of function words
pronoun	Frequency of pronouns
ppron	Frequency of personal pronouns
i	Frequency of 1st person singular
we	Frequency of 1st person plural
you	Frequency of 2nd person
shehe	Frequency of 3rd person singular
they	Frequency of 3rd person plural
ipron	Frequency of impersonal pronouns
article	Frequency of articles
prep	Frequency of prepositions
auxverb	Frequency of auxiliary verbs
adverb	Frequency of common adverbs
conj	Frequency of conjuctions
negate	Frequency of negations
verb	Frequency of regular verbs
adj	Frequency of adjectives
compare	Frequency of comparatives
interrog	Frequency of iterrogatives
number	Frequency of numbers
quant	Frequency of quantifiers
informal	Frequency of informal speech
swear	Frequency of swear words
netspeak	Frequency of netspeak

Table A.2.: Stylistic Metrics *(continued)*

Variable	Description
assent	Frequency of words indicating assent
nonflu	Frequency of nonfluencies
filler	Frequency of illers
AllPunc	Frequency of all punctuation
Period	Frequency of periods
Comma	Frequency of commas
Colon	Frequency of colons
SemiC	Frequency of semicolons
QMark	Frequency of question marks
Exclam	Frequency of exclamation marks
Dash	Frequency of dashes
Quote	Frequency of quotes
Apostro	Frequency of apostrophes
Parenth	Frequency of parentheses (pairs)
OtherP	Frequency of punctuation not captured in above variables
CC	Frequency of coordinating conjunctions
CD	Frequency of cardinal numerals
DT	Frequency of determiners
EX	Frequency of existentials
FW	Frequency of foreign words
IN	Frequency of preposition or subordinating conjunctions
JJ	Frequency of ordinal numbers
JJR	Frequency of comparative adjectives
JJS	Frequency of superlative adjectives
LS	Frequency of list item markers
MD	Frequency of model verbs
NN	Frequency of nouns, singular or mass
NNS	Frequency of plural nouns
NNP	Frequency of singular proper nouns
NNPS	Frequency of plural proper nouns
PDT	Frequency of predeterminers
POS	Frequency of possessive endings

Table A.2.: Stylistic Metrics *(continued)*

Variable	Description
PRP	Frequency of personal pronouns
PRP$	Frequency of possessive pronouns
RB	Frequency of adverbs
RBR	Frequency of comparative adverbs
RBS	Frequency of superlative adverbs
RP	Frequency of particles
SYM	Frequency of symbols
TO	Frequency of to's
UH	Frequency of exclamation/interjections
VB	Frequency of verbs, base form
VBD	Frequency of past tense verbs
VBG	Frequency of present participles
VBN	Frequency of past participles
VBP	Frequency of present tense verbs, other than 3rd person singular
VBZ	Frequency of present tense verbs, 3rd person singular
WDT	Frequency of wh-determiners
WP	Frequency of wh-pronouns
WP$	Frequency of possessive wh-pronouns
WRB	Frequency of wh-adverbs
NER	Frequency of named entities

A.3. Psychological Metrics

Table A.3.: Psychological Metrics

Variable	Description
Analytic	Frequency of words reflecting formal, logical, and hierarchical thinking
Clout	Frequency of words suggesting author is speaking from a position of authority
Authentic	Frequency of words associated with a more honest, personal, and disclosing text
Tone	Frequency of words associated with positive, upbeat style
affect	Frequency of words related to emotions

Table A.3.: Psychological Metrics *(continued)*

Variable	Description
posemo	Frequency of words suggesting positive emotions
negemo	Frequency of words suggesting negative emotions
anx	Frequency of words suggesting anxiety
anger	Frequency of words suggesting anger
sad	Frequency of words suggesting sadness
social	Frequency of social words
family	Frequency of words related to family
friend	Frequency of words related to friends
female	Frequency of female referents
male	Frequency of male referents
cogproc	Frequency of words related to cognitive processes
insight	Frequency of words related to insight
cause	Frequency of words related to causality
discrep	Frequency of words related to discrepancies
tentat	Frequency of words indicating tentativeness
certain	Frequency of words indicating certainty
differ	Frequency of words related to differentiation
percept	Frequency of words related to perceptual processes
see	Frequency of words related to seeing
hear	Frequency of words related to hearing
feel	Frequency of words related to feeling
bio	Frequency of words related to biological processes
body	Frequency of words related to the human body
health	Frequency of words related to health/illness
sexual	Frequency of words related to sexuality
ingest	Frequency of words related to ingesting
drives	Frequency of words related to core drives
affiliation	Frequency of words related to affiliation
achieve	Frequency of words related to achievement
power	Frequency of words related to power
reward	Frequency of words indicating a reward focus
risk	Frequency of words indicating a risk prevention focus

Table A.3.: Psychological Metrics *(continued)*

Variable	Description
focuspast	Frequency of words indicating focus on the past
focuspresent	Frequency of words indicating focus on the present
focusfuture	Frequency of words indicating focus on the future
relativ	Frequency of words related to relativity
motion	Frequency of words related to motion
space	Frequency of words related to space
time	Frequency of words related to time
work	Frequency of words related to work
leisure	Frequency of words related to leisure
home	Frequency of words related to home
money	Frequency of words related to money
relig	Frequency of words related to religion
death	Frequency of words related to death

B. Coefficient Tables

B.1. Body-Level Model

Table B.1.: Coefficient Table for Body-Level Model

Variable	Estimate	Std. Error	z-value	p-value
Intercept	-0.961	0.201	-4.772	< 0.001
iqr_sentence	0.455	0.190	2.396	0.017
iqr_noun_phrase	-0.349	0.194	-1.799	0.072
swc	0.652	0.221	2.945	0.003
types	-0.829	0.381	-2.176	0.03
TTR	0.750	0.317	2.368	0.018
Tone	0.353	0.221	1.596	0.111
function.	-0.174	0.544	-0.321	0.748
we	-0.392	0.310	-1.265	0.206
shehe	-0.038	0.203	-0.189	0.85
they	0.129	0.211	0.608	0.543
prep	0.701	0.345	2.034	0.042
auxverb	0.659	0.371	1.774	0.076
adverb	0.572	0.245	2.334	0.02
negemo	0.432	0.233	1.854	0.064
female	0.530	0.195	2.727	0.006
differ	-0.584	0.260	-2.246	0.025
body	0.696	0.212	3.281	0.001
sexual	-0.527	0.157	-3.354	< 0.001
ingest	0.212	0.178	1.194	0.232
drives	-0.297	0.287	-1.037	0.3
affiliation	-0.548	0.340	-1.613	0.107

Table B.1.: Coefficient Table for Body-Level Model *(continued)*

Variable	Estimate	Std. Error	z-value	p-value
achieve	-0.252	0.199	-1.268	0.205
focuspast	0.042	0.298	0.142	0.887
work	-0.508	0.236	-2.154	0.031
leisure	0.193	0.169	1.139	0.255
money	-0.272	0.204	-1.330	0.184
relig	0.048	0.149	0.322	0.747
AllPunc	-0.432	0.326	-1.322	0.186
Period	-0.615	0.340	-1.808	0.071
Colon	-0.296	0.349	-0.847	0.397
SemiC	-0.870	0.388	-2.245	0.025
Exclam	1.667	0.368	4.527	< 0.001
Quote	0.414	0.183	2.258	0.024
DT	0.614	0.224	2.743	0.006
FW	0.355	0.178	1.995	0.046
JJR	-0.397	0.267	-1.484	0.138
POS	0.070	0.134	0.525	0.6
RBR	-0.396	0.184	-2.151	0.032
RP	0.107	0.144	0.747	0.455
VBD	0.877	0.356	2.463	0.014
VBG	0.577	0.187	3.087	0.002
VBZ	0.125	0.213	0.584	0.559

B.2. Title-Level Model

Table B.2.: Coefficient Table for Title-Level Model

Variable	Estimate	Std. Error	z-value	p-value
(Intercept)	-0.873	0.158	-5.528	< 0.001
types	1.195	0.359	3.327	< 0.001
FOG	-0.790	0.217	-3.641	< 0.001
CL	-1.027	0.225	-4.576	< 0.001

Table B.2.: Coefficient Table for Title-Level Model *(continued)*

Variable	Estimate	Std. Error	z-value	p-value
Tone	0.026	0.214	0.121	0.903
WPS	0.556	0.356	1.560	0.119
Dic	-0.600	0.238	-2.523	0.012
function.	0.618	0.348	1.778	0.075
pronoun	0.143	0.270	0.530	0.596
ppron	0.109	0.209	0.519	0.604
article	-0.605	0.203	-2.975	0.003
prep	-0.101	0.233	-0.433	0.665
adverb	0.215	0.116	1.842	0.065
verb	0.178	0.183	0.975	0.33
adj	0.385	0.125	3.081	0.002
number	-0.388	0.164	-2.375	0.018
posemo	-0.604	0.320	-1.891	0.059
female	0.244	0.150	1.629	0.103
cogproc	-0.179	0.173	-1.037	0.3
cause	-0.242	0.144	-1.685	0.092
see	0.087	0.124	0.699	0.484
hear	-0.944	0.194	-4.879	< 0.001
body	0.398	0.140	2.845	0.004
reward	0.320	0.121	2.647	0.008
risk	-0.361	0.183	-1.973	0.049
focusfuture	0.242	0.128	1.888	0.059
work	-0.206	0.228	-0.903	0.367
money	-0.313	0.134	-2.344	0.019
relig	0.176	0.103	1.711	0.087
death	0.406	0.112	3.639	< 0.001
informal	-0.549	0.214	-2.568	0.01
AllPunc	-0.690	0.229	-3.013	0.003
Comma	-0.545	0.232	-2.345	0.019
Colon	-0.337	0.161	-2.094	0.036
SemiC	-0.279	0.108	-2.578	0.01
Exclam	0.507	0.202	2.511	0.012

Table B.2.: Coefficient Table for Title-Level Model *(continued)*

Variable	Estimate	Std. Error	z-value	p-value
OtherP	0.400	0.131	3.057	0.002
CC	-0.566	0.218	-2.594	0.009
JJR	-0.194	0.192	-1.006	0.314
‘PRP$‘	0.064	0.136	0.470	0.638
TO	0.043	0.119	0.364	0.716
VBD	0.455	0.126	3.610	< 0.001
VBG	0.049	0.121	0.410	0.682
VBN	0.287	0.126	2.277	0.023
VBZ	0.370	0.121	3.051	0.002

Bibliography

[1] O. Ajao, D. Bhowmik, and S. Zargari. Fake news identification on twitter with hybrid CNN and RNN models. *CoRR*, abs/1806.11316, 2018.

[2] D. Avelar. Whatsapp fake news during brazil election 'favoured bolsonaro'. *The Guardian*.

[3] R. Baly, G. Karadzhov, D. Alexandrov, J. R. Glass, and P. Nakov. Predicting factuality of reporting and bias of news media sources. *CoRR*, abs/1810.01765, 2018.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[5] A. Flood. Fake news is 'very real' word of the year for 2017. *The Guardian*.

[6] M. Gruppi, B. D. Horne, and S. Adali. An exploration of unreliable news classification in brazil and the U.S. *CoRR*, abs/1806.02875, 2018.

[7] M. Hervé. *RVAideMemoire: Testing and Plotting Procedures for Biostatistics*, 2020. R package version 0.9-75.

[8] B. D. Horne and S. Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398, 2017.

[9] R. B. J.W. Pennebaker and M. Francis. Linguistic inquiry and word count: Liwc2015.

[10] C. Kang and A. Goldman. In washington pizzeria attack, fake news brought real guns. *New York Times*.

[11] A. Kirilin and M. Strube. Exploting a speaker's credibility to detect fake news. 2018.

[12] M. Kuhn. *caret: Classification and Regression Training*, 2020. R package version 6.0-86.

[13] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[14] N. O'Brien, S. Latessa, G. Evangelopoulos, and X. Boix. The language of fake news: Opening the black-box of deep learning based detectors. In *workshop on "AI for Social Good", NIPS 2018*, Montreal, Canada, 11/2018 2018.

[15] R. Oshikawa, J. Qian, and W. Y. Wang. A survey on natural language processing for fake news detection. *CoRR*, abs/1811.00770, 2018.

[16] M. pui Sally Chan, C. R. Jones, K. H. Jamieson, and D. Albarracín. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546, 2017. PMID: 28895452.

[17] V. L. Rubin, Y. Chen, and N. K. Conroy. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[18] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR*, abs/1809.01286, 2018.

[19] D. Sieradski. B.s. detector.

[20] M. Tosik, A. Mallia, and K. Gangopadhyay. Debunking fake news one feature at a time. *CoRR*, abs/1808.02831, 2018.

[21] S. Tripathi and T. Sharma. Document classification using part of speech in text mining. *International Journal of Science and Research*, 4(12), 2015.

[22] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[23] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *CoRR*, abs/1705.00648, 2017.

[24] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *CoRR*, abs/1905.12616, 2019.